Introduction to the BRC-MH Linux Cluster

22nd July 2010

Introduction

- Name: David To
- BCompSci, MApplSci (Bioinformatics)
- Role: Systems Administrator / Bioinformatician
- Background in:
  - Systems Administration
  - Programming
  - Security
  - Bioinformatics

Funding

- Cluster was purchased by the BRC Nucleus
- £300,000 spent on the cluster
- £75,000 on hosting/support
- NIHR funded, capital grant

Location

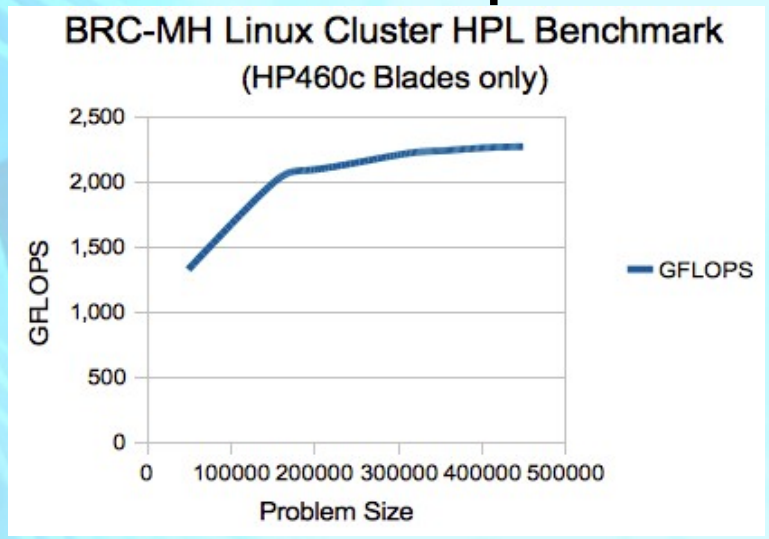
- Currently at Bethlem Royal Hospital
- Will be moved to Denmark Hill this year

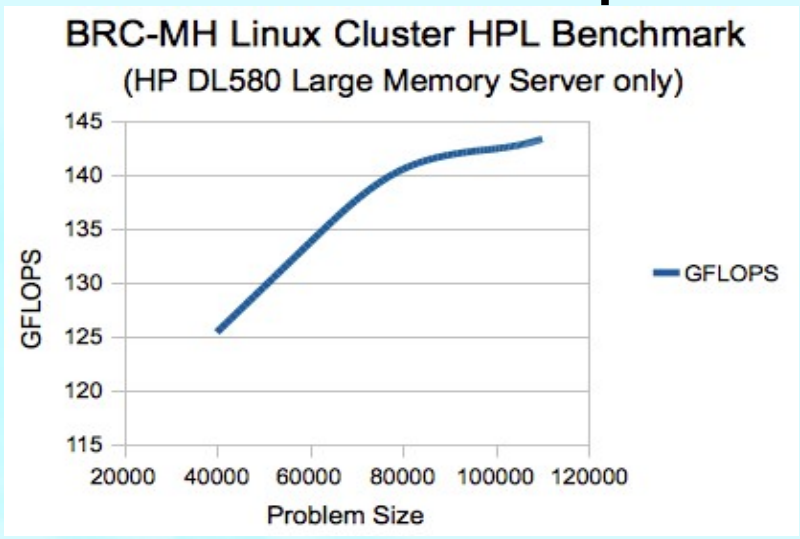
Computational Resources

- 30 x HP 460c Blades
- 1 x DL580
- 264 cores
- 2.087 TB memory

Computational Resources

- Using HPL 2.0
  - 30 Blades: 2.272 TFlops
  - DL580: 142 GFlops
- Combined performance of 2.4 TFlops





Storage

- 120TB of storage provided by 3 Panasas storage arrays.
- Backed up
  - /home
  - /project
- Not backed up
  - /scratch

Backups: Tape

- Daily incremental tape backups
- Essentially everything, but not /scratch
- Maintained by SlaM IT

Backups: Snapshots

- Snapshots allow easy access of old data.
- Hours, daily, weekly
- We currently keep
  - 5 x weekly
  - 7 x of daily
  - 7 x hourly
- Use "`cd .snapshot`" to view snapshots

How to access the cluster

- Access is given via an SSH Gateway
- Authentication is via a Cryptocard two-factor authentication token
- You need to apply for a token to get remote access

Transferring data

- SCP/SFTP
  - Connection to the SSH gateway
  - Outgoing connections from the cluster/gateway not possible (available after Tuesday next week)
- FTP/Web
  - Use wget or curl to download data onto the cluster (goes via a proxy)
- Guest access – not currently possible, but a solution will be available in the near future

Transferring data

- If large amounts of data need to be moved onto/off the cluster use a hard disk for data transport
- We have hard disks for loan to transfer data

Applications

- Generally OS type applications are installed as part of the OS via rpm package management
- Matlab Distributed Computing Server with 64 licenses
- Other less common applications are installed in /share/apps
- Ask if you need applications installed

Perl and R

- Perl CPAN modules
  - /share/cpan – defined in PERL5LIB
- Perl Ensembl modules
  - /share/perl/ensembl/ – defined in PERL5LIB
- R modules (CRAN & Bioconductor)
  - /share/R/library – defined in R_LIBS
- Ask if you need modules installed

Running jobs on the cluster

- Cluster uses Sun Grid Engine for submission of jobs
- 4 Types of jobs
  - Batch
  - Array (parametric)
  - Parallel
  - Interactive

Sun Grid Engine: Batch jobs

- These are single jobs which you would normally run manually on the command line.
- 1 job per submission
- Stdout and stderr is saved to your home directory

Sun Grid Engine: Array jobs

- Like a batch job, but you can specify the number of times it runs, and run on different data, or use different parameters
- Possible to submit thousands of jobs in the one job
- Stdout and stderr is saved to your home directory

Mysql

- Bignode has mysql installed
- Serves out Ensembl databases (read only)
  - homo_sapiens_core_58_37c
  - homo_sapiens_funcgen_58_37c
  - homo_sapiens_cdna_58_37c
  - homo_sapiens_coreexpressionest_58_37c
  - homo_sapiens_coreexpressiongnf_58_37c
  - homo_sapiens_otherfeatures_58_37c
  - homo_sapiens_variation_58_37c
  - homo_sapiens_vega_58_37c
- Access can be granted for your own databases

Questions

- See more at:
- http://compbio.brc.iop.kcl.ac.uk/cluster/
- https://compbio.brc.iop.kcl.ac.uk:8443/biowiki/
- Contact David via:
  - Email: david.to@kcl.ac.uk
  - Skype: brc-mh_linux_support